DOCUMENT RESUME

ED 253 786 CG 017 989

AUTHOR Faley, Robert H.; Kleiman, Lawrence S.
TITLE The Impact of Professional and Administrative Standards on Court Cases Involving Criterion-Related Validity.
PUB DATE Aug 84
NOTE 47p.; Paper presented at the Annual Convention of the American Psychological Association (92nd, Toronto, Ontario, Canada, August 24-27, 1984).
PUB TYPE Information Analyses (070) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Court Litigation; Guidelines; Job Analysis; *Occupational Tests; *Standards; Test Bias; *Test Validity
IDENTIFIERS Adverse Impact; *Criterion Related Validity

ABSTRACT

This paper reviews 12 Title VII court cases litigated since 1978 to assess implications of recent professional and legal guidelines regarding criterion-related validity of paper and pencil tests used by employers to prove job relatedness. Major topics important to an understanding of predictor criterion, including procedural, and data analysis and interpretation issues are examined. Predictor related issues which are considered include choice of predictor, test content, and search for alternatives. Criterion-related issues which are reviewed include criterion relevancy, psychometric considerations, use of the criterion, and training performance as the criterion. Three major procedural issues are examined: (1) choice of predictive or concurrent strategy; (2) sample size; and (3) job grouping. Issues involving magnitude and utility of the predictor-criterion relationship, the use of statistical corrections, cutoff scores, test fairness, and validity generalization are also discussed. Suggestions are offered to guide employers who are interested in successfully conducting or defending a criterion-related validation study. (JAC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED253786

The Impact of Professional and Administrative Standards on Court Cases Involving Criterion-Related Validity

Robert H. Faley*
Krannert Graduate School of Management
Purdue University

Lawrence S. Kleiman
University of Tennessee at Chattanooga

*The authors wish to be considered equal contributers to this manuscript; first authorship was randomly determined. The authors would also like to thank Ann Money and Eddie Pickel for their assistance in the collecting and synthesizing of the legal material that forms a substantial part of this paper. Requests for reprints or inquiries should be sent to Lawrence S. Kleiman, Department of Psychology, University of Tennessee at Chattanooga, Chattanooga, Tennessee 37401.

Running Head: CRITERION-RELATED VALIDITY AND THE COURTS

CG 017989

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Robert H. Faley
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Abstract

Title VII court cases litigated since 1978 were reviewed to assess the implications of the latest professional and legal guidelines for court deliberations regarding the criterion-related validity of paper-and-pencil tests with adverse impact. Major topics important to an understanding of predictor criterion, procedural, and data analysis and interpretation issues were examined. Among the major findings were the heavy reliance placed on test development procedures and the reluctance of many courts to accept recent research findings often contained in professional guidelines which are inconsistent with those in the Uniform Guidelines on Employee Selection Procedures (1978). Suggestions are offered to guide employers who are interested in successfully conducting or defending a criterion-related validation study.

The Implications of Professional and Legal Guidelines for Court Decisions Involving Criterion-Related Validity: A Review and Analysis

When the use of a paper-and-pencil selection instrument results in adverse impact against a "protected group," an employer has the legal responsibility to demonstrate the instrument's job relatedness. Failure to meet this responsibility can have serious financial consequences, especially in class action suits.

This paper seeks to determine how the courts judge the acceptability of criterion-related validation studies when used by employers as evidence of job-relatedness. Such information should provide helpful input for practitioners who must conduct criterion-related validation studies.

While guidelines for performing criterion-related validation studies have been provided by the federal government (Uniform Guidelines, 1978) and by Division 14 of the American Psychological Association (Principles, 1980), a number of practical problems exist with regard to their implementation. First, the legal and professional guidelines are quite detailed, and all requirements can not be met in many cases. From a legal perspective, it would be beneficial to know the minimum standards set by the courts; that is, what the courts consider essential for a favorable ruling. In addition, recent research findings have resulted in a considerable degree of controversy by questioning many of the commonly held assumptions and beliefs concerning criterion-related validity (see Schmidt & Hunter, 1980). As Dreker and Sacket (1983) note "selection research represents, in actuality, a dynamic and complex process with many controversies and questions left unanswered" (p. 4). Since research findings and legal and professional guidelines are sometimes inconsistent, it is

important for validation specialists to know how the courts have typically resolved these inconsistencies.

The scope of this paper is limited to a review and analysis of Title VII cases since 1978 in which evidence of criterion-related validity was used to establish the job-relatedness of a paper-and-pencil selection test that had adverse impact. Only cases adjudicated since 1978 are reviewed in order to better assess the impact of the latest legal and professional guidelines on the courts' deliberations. In all, 12 cases which met the above criteria were identified by means of a computer search (Lexis Nexis), which contains all published and unpublished cases.

Predictor-Related Issues

Choice of Predictor

According to Principles, when a criterion-related strategy is used to validate a selection instrument, the choice of the predictor should be based on an empirical, logical, or theoretical foundation; i.e., there should be some sound reason for believing it to be job-related. However, serendipitous findings are acceptable if appropriately verified.

The courts were divided on the importance of the need for a "rationally based" choice of predictor instruments. Two separate approaches emerged. In half of the cases, the courts were unconcerned about how the selection instrument was developed. Judgments of job-relatedness were based solely on evidence pertaining to the validation studies. These courts apparently believe the "proof is in the pudding," and focus their deliberations on "outcome" rather than "process" information. Serendipitous findings are acceptable to these courts.

In the remaining cases, however, test development procedures (i.e.,

"process" information) had a direct bearing on the courts' decisions. Foremost, these courts looked to see whether the development or choice of the test was based on an adequate job analysis. Where a job analysis was performed (viz., Bridgeport, Rivera, and Walls), its adequacy was carefully scrutinized.

In Bridgeport, for example, the court ruled against the defendants because of the "casual nature" of their job analysis. Specifically, the court determined that the defendant's job analysis was deficient because the "list of job qualities" derived from the analysis were too broadly defined, and because the analysis failed to weight these job qualities according to their importance. The Walls court also criticized the "broadness" of the task and KSA definitions derived from the defendant's job analysis (e.g., "knowledge of general sociological aspects"), and the fact that the worker attributes measured by many of the subtests were linked to unimportant task behaviors (i.e., tasks judged to be unnecessary for sucessful performance at entry). Only in Rivera was the court satisfied that the job analysis was comprehensive, and that the worker attributes measured by the examination were critical requirements for successful performance. Unfortunately, neither the job analysis nor the test development procedures were described in any detail in the written decision.

Where dissatisfied with the manner in which an organization chose or developed its selection instrument, many courts ruled that the defendant's burden to substantiate the validity of the predictor was considerably greater than it would have been had the choice been more rational or systematic (e.g., had it been based on a job analysis). With one exception (Cormier), the court ruled for the plaintiff in each of the cases in which the defendants' job analysis was inadequate, or non-existent. This even occurred in Bridgeport, where exam scores were significantly correlated with measures of job

performance! Where satisfied with the job analysis (e.g., Rivera), the court ruled for the defendant.

The rationale of some courts for placing such great importance on this issue is summed up best in Guardians:

> Taking the quality of test preparation into account in judging test validity would be useful in lessening the risk that a court would fall into error in umpiring a battle of experts who speak a language it does not fully understand (p. 243),

and in I.M.A.G.E.:

> Insufficient spadework usually results in a poor garden (p. 808).

The implication of such "process-oriented" court decisions for employers is clear--any selection instrument must be based on a thorough job analysis as well as designed to measure all the important attributes an employee should possess at entry into a job. In the minds of courts that make their rulings based on developmental issues, a systematic approach to test development based on a thorough job analysis increases substantially the likelihood that the test will result in valid predictions of job performance. Moreover, while not required by them, such an approach should also satisfy courts more interested in validity results than developmental procedures.

Employers, however, are often faced with having to validate a test currently in use, which has not been developed on the basis of a systematic job analysis. The probability of successfully defending such an exam would be quite low if tried in a court which places a heavy emphasis on test development issues. An employer faced with this set of circumstances should either perform

a "post hoc" job analysis prior to validation (which, hopefully, would establish the relationship of the test to important job attributes) or develop a new test (based on a thorough job analysis) to be used in lieu of the existing one. A third choice would be to "go forward" with evidence of the test's criterion-related validity, but only if that evidence is "substantial."

Test Content

Test content issues deal with the courts' preferences regarding the predictor's "type" (e.g., whether it is an achievement or aptitude test, work sample, or interest or personality inventory, etc.) as well the nature of the questions on the predictor (e.g., multiple choice, essay, etc.). Unlike the restrictions placed on the type of predictor when a content validation strategy is employed (e.g., no intelligence tests), the Uniform Guidelines place no such restrictions when a criterion-related strategy is used.

Although most courts did not scrutinize test content, there were four cases in which test content issues were addressed. In two of these (I.M.A.G.E. and Dickerson), the court indirectly criticized the appropriateness of examination type. In I.M.A.G.E., for instance, the examination consisted of three parts: address-checking, oral directions and vocabulary. The court, apparently swayed by testimony that such tests are generally invalid, and that most employers have eliminated these types of subtests since the Grigg's decision, ruled that they were inappropriate. In Dickerson, the court commented favorably on the fact that the defendant had stopped using the Wonderlic exam since "a number of cases since Griggs have resulted in the discontinuance of the Wonderlic exam because of unvalidated use with disparate impact" (p. 1330). On the other hand, the Cormier court upheld the use of the Wonderlic, which in this instance was supported by validity evidence.

It thus appears that many courts are unwilling to accept certain tests (e.g., Wonderlic) or certain types of tests (e.g., vocabulary) simply because other courts have ruled against them. This is unfortunate, since these earlier adverse decisions are often the result of the fact that the tests were never validated (e.g., Griggs).

In the remaining two cases in which the courts addressed test content issues, the courts were concerned about the type and content of the test items. For example, in Bridgeport, the court criticized the defendant's choice of item type. Here the court was skeptical of the defendant's use of a multiple-choice format.

> While this type of test [multiple choice] has demonstrated some ability to 'predict' subsequent performance of individuals on other written exams, it does not follow that multiple-choice exams are adequate to predict all types of human behavior. In fact,...these tests are inadequate to the task of predicting complex patterns of behavior (p. 758).

Alternatively, the Cormier court, in response to a criticism by the plaintiff that the content of the test items was not job-related, stated:

> It is inappropriate for the layman to base conclusions on the appropriateness of tests like the Wonderlic or Bennett based on their own perception of the items of the test (p. 690).

The scrutinization of test content appears to be another example of the desire of some courts to focus their attention on non-empirical issues or to

sidestep empirical issues before them. Rather than allowing evidence of criterion-related validity to determine whether the test before them adequately predicts job performance (as in Cormier), these courts prefer to pre-judge the validity of the examination based on an assessment of the content of the test. Ironically, Schmidt, Hunter, Croll and McKenzie (1983) have recently demonstrated that judgments of a test's validity based on an examination of its content can be quite accurate, if performed by experts in the testing field.

Fortunately for employers, the issue of test content was not a major one. However, in light of the decisions in I.M.A.G.E. and Dickerson, employers should be wary of the use of general aptitude tests. This may be an unfortunate circumstance given the recent positive findings concerning the validity of such tests for jobs with a high cognitive component (Gutenberg, Arvey, Osburn, and Jeanneret, 1983).

Search for Alternatives

The Uniform Guidelines specify that an employer should, as part of a validity study, investigate suitable alternatives which have as little adverse impact as possible. Of the two or more procedures with essentially equal validity, the one with the least adverse impact should be used.

The issue of an alternatives search was addressed in six cases. In none of them did the employer "formally" search for a suitable alternative. However, in two cases (Pegues and Cormier), the courts ruled that the employers had nonetheless met their burden. In Pegues, the employer used a number of General Aptitude Battery Tests as predictors of job performance. The "alternatives" requirement was satisfied by testimony indicating that "an individual test with a high validity coefficient may not be used in the final battery if another test produces better results for a minority subsample" (p. 254). In Cormier, the

defendant satisfied the requirement by lowering the cutoff scores on the examinations.

The remaining courts addressing this issue reached the conclusion that the burden of searching for alternatives was either unnecessary (O'Neill), or should be placed on the plaintiff (Rivera, Dickerson, and I.M.A.G.E.). The O'Neill court determined that it was not within its jurisdiction to force an employer to consider alternatives:

> It is equally beyond the province of a federal district court to control management decisions concerning the acceptability of risks involved in alternative methods of selection for promotion within the Police Department, so long as such decisions do not serve to mask racial discrimination (p. 465).

The rationale for disregarding the Uniform Guidelines and placing the burden of search on the plaintiff was discussed in I.M.A.G.E. This court stated that it would not enforce the requirements of the Uniform Guidelines "in the face of the case law distributing the burden of proof differently" (p. 812).

In Cormier, the defendant had met its burden by lowering cutoff scores, thus reducing adverse impact. Nonetheless, the plaintiff asked the court to consider the suitability of three additional alternatives to the tests used by the defendant. These were biodata, work sample tests, and trainability tests. This attempt to influence the court failed for two reasons. First, the plaintiff failed to prove that these alternatives were indeed suitable. That is, the plaintiff could not demonstrate the existence of another specific test which had been validated on jobs similar to those in the instant case that met

the "alternatives" criteria. Second, the court rejected the notion that the plaintiff had the right to suggest additional alternatives when the employer had already implemented one (viz., lowered cutoff scores). According to the court, the "plaintiff's approach suggests a meaningless and endless cosmic search for possible alternatives" (p. 699).

The alternatives search issue apparently is not an overly important one to the courts. It is rarely addressed, and when it is, the courts have most often held the plaintiff responsible for the "search." This lack of court concern may be somewhat justifiable, given the evidence reported by Reilly and Chao (1982) that only biodata and peer ratings appear to have validites equivalent to those of written tests. Moreover, evidence that these alternatives also have less adverse impact than traditional paper-and-pencil tests is unavailable.

Even if an employer is freed (because of judicial precedent) from the responsibility of such a search, the issue would still be salient if the plaintiff could demonstrate the existence of a suitable alternative.

A number of unresolved issues would pertain to this type of situation. One issue hinges on a comparison of the adverse impact of each instrument. How much "less" adverse impact should the alternative have in order to be considered suitable? Would any difference in adverse impact be acceptable, or should the difference be statistically significant? The second issue deals with how the court should decide whether the validity of the alternative is "at least equal" to that of the original instrument? Given the lack of statistical power found in most validation studies (Schmidt, Hunter and Urry, 1976), the difference between the two observed validity coefficients would have to be quite sizable in order to reject the null hypothesis. Conceivably, the situation could arise in which a nonvalid selection device is judged to be a suitable alternative to a

valid one, simply because the two observed validity coefficients do not significantly differ.

Another issue regarding the search for a suitable alternative concerns the concept of utility. Should an employer be forced to use an alternative which is "suitable" (i.e., compared to the original instrument, has less adverse impact and equal validity) yet has less utility (e.g., has a considerably greater cost)? For example, it is conceivable that a court could require an employer to use an assessment center if it were shown to be a suitable alternative, even though the cost of installing and using such a procedure would be prohibitive.[1]

Criterion-Related Issues

Criterion Relevancy

A principle requirement of any criterion is the relevance of its content. The Uniform Guidelines state that criteria are relevant to the extent that they represent important job duties, work behaviors, or work outcomes as developed from a review of job information (i.e., a job analysis).

Importance of job analysis information. In most cases the courts' judgments of criterion relevancy were based on the presence of a properly conducted job analysis. For example, the court ruled against the criterion measures used in Dickerson, Chicago, Bridgeport and one of the measures used in Ensley Branch because of the failure to base them on information derived from an adequate job analysis. In Bridgeport, the court stated that "the insufficient data collected...concerning the job characteristics of a firefighter tainted all subsequent efforts to develop a job-related exam" (p. 756). And in Dickerson, the court concluded that "this failure of the criteria, which by itself would render the studies meaningless, is another element in the court's overall rejection of the studies" (p. 1341).

In two cases the criterion measures were acceptable (i.e., judged to be relevant) because they were developed based on information from an adequate job analysis. The O'Neill court, for instance, approved the criterion development procedures conducted in two separate validation studies. In one study the criterion consisted of six dimensions derived from a factor analysis of 91 job analysis task statements; the other consisted of 10 performance criteria which were based on a job analysis and consultation with command personnel.

There were, however, two cases in which the criterion measures were judged to be relevant in the absence of a job analysis (Pegues and Cormier). In Pegues, the criterion consisted of ratings of quality, quantity, accuracy, job knowledge and versatility. The court disagreed with the plaintiff's claim that such criteria were irrelevant, and ruled that they were central aspects of job performance. In Cormier, the court concluded that a criterion instrument which included 10 "universal" criteria (e.g., understanding instructions, handling on-the-job problems, job learning ability) were acceptable. The court made its judgment based on expert testimony and its "familiarity with the jobs in question developed through the testimony of other witnesses" (p. 691).

Psychometric considerations. Factors in addition to whether the criterion was based on an adequate job analysis were sometimes considered in assessing the relevancy of the criterion. These included, among others, criterion contamination, dimensional definitions, and weighting and scale type. When criterion contamination was the standard for determining the extent of criterion relevancy, the courts looked to see whether the actual or operational criterion included variance that was unrelated to the conceptual criterion. For example, the effects of the interrelated performance of other workers on individual performance was considered by the Chicago court to be a contaminating influence

on individual ratings of performance. Moreover, the criterion used in Chicago assessed performance on the employees' present job, rather than the one for which selection was being made. This practice was unacceptable to the court because the defendant offered no evidence that the requirements of the two jobs were identical.

The Uniform Guidelines and Principles also suggest that the operational criterion should be unambiguous. When ratings of performance are used as the source of criterion information this can be accomplished by either improving the definitions of the performance dimensions and/or by providing anchors for the various rating scale points. All cases involved the use of ratings to gather job performance information.

In two cases (Bridgeport and Dickerson), the courts rejected the criterion measures because the performance dimensions they attempted to assess were vaguely defined. The question of dimensional definitions and scale properties also arose in Pegues. Here, the plaintiff questioned both the clarity of the dimensional definitions and lack of specific behavioral anchors of the rating scales. The court, however, considered the performance dimensions to be "self explanatory" and the use of specific anchors unnecessary in light of research evidence presented by the defendant's expert witness that anchored scales are not superior to unanchored ones.

This is surprising. Based on a review of the research conducted during the past 30 years, Landy and Farr (1980) concluded that, while rating scale format itself is not particularly important, it is crucial that careful attention be paid to item selection and anchoring procedures. According to these authors, behavioral anchors were preferable to simple numerical or adjectival anchors, especially in the absence of good dimension definitions.

Specific questions about the type of scale also arose in Bridgeport. This court criticized the defendant's use of a ranking rather than a rating procedure to assess job performance. Supervisors were instructed to rank each of their subordinates from best to worst. This procedure was criticized because "it assumes that the best firefighter under one supervisor is better than the second best firefighter under another supervisor" (p. 756).

Another issue related to criterion relevancy deals with the weighting of criterion dimensions. According to the Uniform Guidelines and Principles, measures of the criticality or importance of work behaviors to be assessed on the criterion instrument are essential to a credible criterion-related study. This issue was addressed in only one case (O'Neill). Here, the court was satisfied with the defendant's system of weighting each performance dimension according to its relative importance (based on a combined weight-index the defendant developed). It did not, however, scrutinize the weight index itself. Surprisingly, failure of employers to systematically weight criterion components went unnoticed by the courts in the remaining cases.

The final psychometric issue involves the courts' perception of measurement error in the criterion. Unlike most courts which viewed measurement error in a negative light, the Cormier court noted that its presence served to lower the observed validity coefficient. The court concluded that a significant validity coefficient, despite a poor criterion measure, merely strengthened the defendant's case.

Use of the criterion. Another set of criterion-related issues the courts addressed dealt with criterion usage (i.e., the rating process). In terms of usage, the courts were mainly interested in the preparedness of the raters to provide accurate and unbiased ratings (e.g., whether rates were given clear

instructions, an opportunity to observe and discuss the jobholder's performance, and/or training).

The inclusion of adequate instructions was an important issue in three cases (Cormier, O'Neill, and Dickerson). The Cormier and O'Neill courts were satisfied with the instructions given the raters. In O'Neill, the instructions included an explanation of how to complete the rating forms, and a discussion of the need for rater impartiality and objectivity. The Cormier court was impressed with the fact that the defendant's consultant personally gathered the performance ratings. "He gave instructions to each rater and satisfied himself that the rater understood and could properly fill out the rating form" (p. 691). The inclusion of instructions, however, did not satisfy the Dickerson court. Here, the raters were instructed to avoid leniency. Despite these instructions the ratings were "high." The court considered this as evidence that the raters disregarded the instructions.

The rater's opportunity to observe the ratee's performance and to discuss the performance with another rater were important issues in Walls and Dickerson. The Dickerson court criticized the unavailability of a second rater with whom the primary rater could have conferred. And in Walls, the court criticized the fact that the raters had insufficient opportunity to observe the ratees' performance.

In none of the cases did the defendants provide evidence of any rater training. This failure was criticized in three cases (Bridgeport, Dickerson, and Walls). The Pegues court, however, was satisfied with the raters' capabilities despite evidence of no formal training because, as supervisors, they "probably" had considerable experience in assigning ratings.

Training performance as the criterion. The final criterion issue the

courts have examined involves the appropriateness of using training performance as a criterion measure. According to the Uniform Guidelines and Principles, success in training is an appropriate criterion when it is both shown to be related to successful job performance and is properly measured.

Court decisions have been mixed on this issue, depending on each court's interpretation of the relevance of the 1976 Washington v. Davis Supreme Court decision. For example, in Virginia and Rivera, the court interpreted Washington v. Davis as not requiring evidence that success in training is related to successful job performance. However, in Ensley Branch and Guardians, the court concluded that Washington v. Davis was not applicable to Title VII cases, and ruled against the employers because they failed to demonstrate a relationship between training and job performance. Finally, in Dickerson, the court ruled that training performance was not an adequate criterion measure based on the fact that training performance was inaccurately measured.

Other issues related to criterion development and measurement that deserve attention, but were not addressed (or addressed in enough detail) by the courts include: criterion reliability, the use of multiple versus composite criteria, attention to rater selection (and the use of multiple raters), and common method variance between test and criterion.

No single case addressed all the factors which the Uniform Guidelines and Principles suggest are important for determining the adequacy of a criterion measure. Three factors emerged as most critical from the courts' point of view. In order of importance, these are job analysis, rater-preparedness, and the use of training as a criterion.

It is not surprising that the presence of a thorough job analysis appears to be the primary issue in the courts' determinations of criterion relevancy.

In this regard, the courts' view is consistent with the requirements stated in the Uniform Guidelines and Principles.

The major implication of the great weight placed on a job analysis by the courts pertains to those employers who use information from existing performance appraisal systems as the source of criterion data, especially where the appraisal instrument has not been developed based on an analysis of the job(s) for which the test is validated. If not based directly on an job analysis of these job(s), a "special" performance appraisal instrument should be developed for the purpose of the validation study in order to comply with court requirements.

In terms of criterion usage, the courts apparently agree with the idea that clear instructions and intensive rater training that focuses on how to observe and record work-related behaviors systematically and representatively, may well be the key to making accurate, error-free appraisals of employee performance. Employers would be well advised to review the rather extensive research on rater training (e.g., Spool, 1978), and based on that information, develop or purchase a suitable training program.

Recommendations regarding the use of training performance as a criterion are difficult to make since the courts are divided on this issue. The safest approach would be for employers to avoid the use of training performance as a validation criterion unless they can show a significant training/job performance relationship, and also that training performance itself is properly measured.

Procedural Issues

Three major procedural issues pertaining to criterion-related validation were examined by the courts--choice of predictive or concurrent strategy, sample size, and job grouping.

Choice of Validation Strategy

Both the Uniform Guidelines and Principles state a preference for a predictive over a concurrent validation strategy, but also recommend the use of the latter where a predictive study is not feasible. Based on an analysis of several research studies (e.g., see Bemis, 1968 or Pearlman, Schmidt and Hunter, 1980), the Principles clearly indicate that well conducted concurrent studies can provide useful estimates of predictive validity.

Cases that involved the validation of the predictor against a measure of job performance all employed a concurrent strategy. Those that involved validation against a training performance criterion employed a predictive strategy.

While most courts stated a preference for a predictive strategy regardless of the type of criterion measure, concurrent validation was generally acceptable to the courts. However, the courts did not unquestionably accept the use of a concurrent strategy by defendants. The courts' primary considerations regarding the appropriateness of the concurrent strategy revolved around two issues--the biasing effects of job experience and the representativeness of the validation sample with regard to the percentage of protected group members. Obviously, the courts believe that validity coefficients become more meaningful the more closely the group on which the coefficient is based resembles the group on which the test will eventually be used.

For example, the Walls court criticized the use of a concurrent design because it felt job experience would bias test performance results. The Dickerson court echoed these concerns, stating a concurrent design was clearly inappropriate for knowledge tests or where it could be demonstrated that on-the-job experience would affect an applicant's test score. However, this

court ruled the test was valid even though a concurrent validation strategy was employed, because the plaintiff failed to show that job experience did, in fact, bias test score results.

On the issue of sample representativeness, the courts have differed both in terms of whose responsibility they feel it is to show the validation sample is a representative one, and on the importance of representativeness considerations themselves. For example, in Dickerson, the court was very concerned with the representativeness of the validity sample and placed the burden on the defendant to show that it was racially balanced. In Pegues, the burden was placed on the plaintiff to show it wasn't. The Pegues court further stated that the issue of representativeness was of little importance, citing expert testimony that research showed that differences, for example, in applicant racial characteristics do not moderate or change validity (i.e., that differential validity is a "pseudophenomenon").

Defendants who decide to employ a concurrent validation strategy must be prepared to demonstrate that the resultant validity coefficient can be generalized to the applicant population. This would best be accomplished by matching, if possible, the validation sample with the applicant pool on factors such as education, experience, sex, race, etc., or by using statistical techniques (such as partial correlation or analysis of covariance) to remove from validity results any unwarranted influence due to these factors.

Plaintiffs attacking validity results based on the use of a concurrent strategy, on the other hand, should investigate whether significant differences in test performance occur between incumbents and applicants. Moreover, if this information is not available, plaintiffs should investigate whether significant differences between incumbents and applicants occur on any of the factors

suggested.

Sample Size

Sample size is an important procedural issue because it influences the statistical power of the test/criterion relationship. The sample size necessary to produce adequate power is much larger than has typically been assumed (Schmidt et al., 1976; Sackett & Wade, 1983); in cases where sample size is deficient there is a greater likelihood that a "true" significant validity coefficient would not be detected. The Uniform Guidelines and Principles recommend that, to be acceptable, the validation sample must be large enough to provide adequate statistical power. "A study which has a low probability of detecting the true validity of the predictor provides little information" (Principles, p. 9).

In only one case reviewed (Cormier) was sample size and its implications for statistical power discussed by the courts. Here, the sample size (n = 20) was "too small for appropriate analysis" (p. 693). The court ruled, however, that the test was valid despite the absence of a statistically significant validity coefficient, based on the opinion of the defendant's expert witness that because of the small sample size a type II error had occurred. One would think that defendant's whose validity coefficients were marginally significant and low, and where sample sizes were less than several hundred, would "jump" to introduce similar evidence to explain the lack of statistical significance of their validity coefficients.

Job Grouping

Job grouping was the final procedural issue addressed by the courts. When the same test was used to select applicants into jobs with different titles (e.g., police sergeant and lieutenant), the courts required a justification of

grouping procedures. Justification has invariably been based on the courts' subjective judgments of the comparability of the jobs in question. For example, in Dickerson the selection device was used for entrance into a variety of apprenticeship training programs, which were lumped together for the purpose of validation. The court criticized this pooling of jobs, stating:

> All apprentices must have classroom knowledge, on-the-job knowledge, and specific trade skills, but the content of these three elements varies significantly. The Court interprets the Uniform Guidelines as requiring content or skill similarity in order to combine jobs for validation (p. 1336).

On the other hand, the courts appear to be unconcerned about groupings within job title (e.g., "desk" versus "field" sergeant). This is somewhat at odds with the Uniform Guidelines and Principles, which require that any job grouping be based on a showing that the different jobs require substantially the same work behaviors or attributes. Obviously, job "title" should not be the sole determinant of whether or not jobs treated as the same for validation purposes are clearly similar or different.

It is surprising that defendants have attempted to use statistical techniques to help justify job groupings in only one instance. In Cormier, statistical evidence based on data derived from job analysis information was employed to justify job groupings. Four job clusters were identified through the use of a hierarchical clustering technique; the tests were validated separately for each cluster.

However, although statistical techniques can provide useful evidence to

support job groupings for validation purposes, Pearlman (1980) notes that they cannot be used as the sole input because they can result in groupings that vary as a function of the grouping technique, job descriptions, and/or the objective for which the job family has been designed. Because of these problems with statistical grouping techniques, the question of grouping jobs is still principally a subjective, not a statistical one. According to Cascio (1982), if statistical grouping techniques are used, adequate job analysis information plus replication of results in two or more independent samples are essential.

Data Analysis and Interpretation

Magnitude and Utility of the Predictor-Criterion Relationship

Other things being equal, a selection instrument is "appropriate" for use to the extent that it correlates highly with one or more important aspects of job performance (Uniform Guidelines). Normally, the correlation must be statistically significant. Departure from this practice should be based on reasons (e.g., utility) which must be stated in advance (Principles). The Uniform Guidelines further state that when supported by evidence of a statistically significant but low validity coefficient, the test should be more carefully scrutinized with regard to adverse impact and to the availability of alternative selection procedures which have less adverse impact.

The utility of a selection device refers to the degree its use improves the quality of individuals selected, compared to that which would have occurred had it not been used. The most commonly employed methods of utility analysis are Taylor-Russel tables, Naylor-Shine tables, and the Brogden and Cronbach-Gleser models.

The courts took one of two approaches in determining the "appropriateness" of the predictor-criterion relationship. In some instances, appropriateness was

based solely on the magnitude and/or significance of the validity coefficient. If the coefficient was significant or suitably high, it was considered acceptable. This approach was taken by the I.M.A.G.E., Bridgeport, Rivera, and Guardians courts. The validity coefficients were judged acceptable by the latter three courts because they were statistically significant. Only in Rivera was the actual coefficient reported (r = .38). The I.M.A.G.E. court ruled that the coefficients were too low to be of any practical use. None of the comparisons showed even a "marginally acceptable" positive correlation.

The other approach that was used to determine the adequacy of the test-criterion relationship was to jointly consider both the validity coefficient and the utility of the examination. This was especially true in those instances where the magnitude of the defendant's validity coefficient was low, yet significant. In three of the cases in which this joint approach was taken, the courts ruled that the test-criterion relationships were inadequate because the coefficients were low and/or non-significant and the tests had little, if any, utility. In Dickerson, the test-criterion correlations were each significant, but of low magnitude, ranging from .22 to .27. The court stated that while low magnitude is not, per se, evidence of low utility, it does trigger a closer scrutiny. Specifically, the court considered five separate indicators of utility: (1) the importance and number of job performance components well predicted by the test, (2) the selection ratio, (3) base rate versus the success rate in using the selection device, (4) the risk factor associated with the selection of false positives, and (5) costs.

The Dickerson court concluded that the test-criterion relationship was inadequate based primarily on the fact that the one component of job performance which was well predicted was relatively unimportant. Also influencing the

decision was the low coefficient of determination (which was preferred over the Brogden model as a method of utility analysis), and the fact that the success rate using the exam was "inappropriately" compared to the expected success rate associated with random selection. The court viewed this comparison as unrealistic.

In Walls, the validity coefficients were low, and only a few were statistically significant. The court ruled against the tests on the basis of evidence demonstrating that the number of successful performers selected by the tests were only two percentage points greater than the number that would have been chosen by random selection. The court was also influenced by the fact that forty-seven percent of those failing the tests could have succeeded on the job.

The Ensley Branch court also ruled against the adequacy of the predictor-criterion relationship. The validity coefficient, while statistically significant, was low, as was the utility. The utility of the test was judged both on the basis of the standard error of estimate for predicting job performance and the degree of adverse impact.

In four other instances in which the courts considered both validity and utility evidence jointly, the predictor-criterion relationships were considered acceptable. In Virginia, the test-training performance correlations were .42 and .55, both highly significant. The court ruled that in addition to statistical significance, the tests had practical significance, as evidenced by the sharp increase in attrition rates during training since the tests were dropped.

In Pegues, the reported validities were low, but the court considered the tests to have practical value because of the considerable dollar value associated with the reported levels of validity and the grave consequences

associated with false positives. In O'Neill, while low validity coefficients were reported, the tests were nonetheless judged useful since the selection ratio was low and the risk value associated with selecting applicants who were unqualified, as in Pegues, was high. Finally, in Cormier, the validity coefficients of the examinations chosen for the two job clusters judged to have a large enough sample size were .27 and .47. In addition to these significant validity coefficients, the defendants demonstrated (using Brogden's model) that the use of these exams enabled them to gain several hundred thousand dollars a year in increased productivity.

Even though defendants can receive a favorable ruling based on evidence of a significant and large validity coefficient, validity coefficients with large magnitudes are not the rule in most studies (Ghiselli, 1973). Considering the fact that power is generally low in most validation studies because of small sample sizes, Type II error is high.

It would appear, then, that it is in the best interests of defendants to present strong utility evidence if validation sample size is small and the magnitude of the validity coefficient is small and significant. Even a marginally significant validity coefficient can receive credible support in the eyes of the courts if utility evidence is good. Under similar circumstances, even a nonsignificant validity coefficient could possibly be defended as the product of Type II error. Reliance on validity, without supportive utility evidence, can have adverse consequences for many defendants.

The Use of Statistical Corrections

On the issue of statistical corrections, court determinations appear fairly uniform. Although the use of statistical adjustments is recommended in Principles and addressed in the Uniform Guidelines, the courts are usually

reluctant to accept them.

Principles recommend that corrections for criterion unreliability and predictor range restriction should be made whenever the necessary information can be obtained. Although not as positive, the Uniform Guidelines suggest that if statistical adjustments are made they should be thoroughly described and explained. Both the Principles and Uniform Guidelines require that uncorrected validity coefficients also be reported.

The only statistical correction addressed in the cases reviewed concerned range restriction in the predictor. Correction for attenuation due to criterion unreliability was not addressed. The courts' objections to statistical corrections rested on two concerns. First, corrections were perceived by the courts as potentially misleading. That is, they were viewed as a means by which defendants could "artificially" increase the magnitude of their validity coefficients. Second, some counts did not agree with the underlying normality assumptions associated with the use of statistical corrections. For example, both the Dickerson and Ensley Branch courts ruled against the use of range restriction corrections because they felt the normality assumption underlying the predictor scores was just that--an assumption. Neither defendant was able to produce sufficient evidence to persuade the courts that this assumption was valid. Although the use of a corrected validity coefficient was sanctioned in O'Neill, the court did not indicate its rationale for accepting the correction.

Unless defendants can better explain to the courts the rationale for using statistical corrections, they are better off avoiding them (or at least, not depending on them), even though professional guidelines recommend their use. Apparently, the use of statistical corrections fosters the perception in the courts' minds that employers are artificially tampering with validity evidence

to make it appear better than it "really" is. We recommend that employers provide the courts with evidence of the potential viability of such corrections. For instance, employers could introduce as evidence the recent finding that statistical corrections for attenuation and range restriction generally provide underestimates of the "true" validity under most circumstances (Bobko, 1983).

Cutoff Scores

The issue of predictor cutoff scores represents an area of measurement where the rationale for the choice of pass/fail determinations is paramount. Both the Uniform Guidelines and Principles recommend that cutoff scores should be based on considerations related to utility, including social policy and adverse impact. A number of statistical methods have been developed to help make the determination of optimum cutoff scores more objective (e.g., see Darlington and Stauffer, 1966 or Rorer, Hoffman, LaForge, and Hsieh, 1966). Unfortunately, they all have underlying assumptions (e.g., all incorrect classifications are equally costly) that make it difficult to apply them in most practical selection situations. Thus, the selection of cutoff scores is typically based on what is perceived as a "reasonable standard."

The proper cutoff in most cases reviewed rested mainly on considerations of the degree of adverse impact. For example, in I.M.A.G.E. and O'Neill, the court supported cutoff scores selected to minimize adverse impact, regardless of other utility considerations. Moreover, in Bridgeport, the court ruled against an arbitrarily set cutoff score, stating it represented an "unsound approach" that resulted in increased adverse impact. Alternatively, the use of economic considerations and the degree of risk associated with an incorrect selection decision were considered justifiable reasons for setting the cutoff score in Virginia, even though this resulted in a large degree of adverse impact.[2]

Although considerations regarding adverse impact appear to weigh heavily on the minds of the courts as they scrutinize cutoff score decisions, the overriding factor appears to be the credibility of the evidence used to support cutoff determinations. An informally or arbitrarily determined cutoff score will not be tolerated by most courts (unless, possibly, adverse impact is minimal). Formalized procedures that take into account cost/benefit, selection, and success ratios as well as other important factors, have a greater chance of passing the courts' scrutiny, even where the adverse impact that might result is high. Thus, defendants will best be able to support cutoff scores where the procedure for determining the cutoff is formalized, utility is high, the job involves safety considerations (e.g., police or firefighter), and adverse impact is low. Even when adverse impact is high, if the first three considerations are met, a ruling favorable to the defendant could result.

Test Fairness

The Uniform Guidelines state that unfairness occurs when members of one protected group score lower on a selection test than members of another, and differences in test scores do not correspond to group differences in job performance. The Uniform Guidelines require the user to investigate unfairness if technically feasible (i.e., where minority group subsamples are large enough). When it is not feasible to conduct a fairness study the user may rely on the weight of evidence from other validity studies of the same or similar jobs.

The feasibility of test fairness studies was addressed in three cases. In Rivera, the court concluded that a test fairness study was infeasible because of the small minority sample size (n was not reported). In Walls, three validation studies were examined. The minority sample sizes were 32, 37 and 78. The court

ruled that feasibility existed in the latter study only. Three studies were also reported in Dickerson. The court used "30" as the standard to determine whether the number of minorities was sufficient. This standard was met in only one of the studies (minority $n = 42$). The defendant, however, attempted to demonstrate the fairness of the tests in the other two studies by providing evidence of fairness gathered on the tests in another setting. The Dickerson court did not admit this evidence since the two settings were not comparable (cf., Pegues).

Evidence of test fairness has been presented to the courts using a variety of approaches (e.g., single-group validity, differential validity, and various models of test fairness which determine whether persons with equal probabilities of success on the job have equal probabilities of being hired). A number of researchers have suggested that the investigation of single-group and differential validity is of little utility to an organization (e.g., Hunter, Schmidt, & Hunter, 1979). Two reasons have been advanced. First, these approaches are conceptually incompatible with the notion of test fairness; mean job performance must also be considered. Second, investigations of single-group and differential validity suggest the occurrence of either is rare (Boehm, 1977; Hunter, Schmidt, & Hunter, 1979). Perhaps the most widely professionally accepted model for determining test fairness is the Cleary Model, which tests regression line and intercept differences between subgroups.

In two cases the court focused on the issue of subgroup differences on test scores to assess fairness. The O'Neill court criticized the defendant for not performing a differential item analysis which could have identified and eliminated discriminatory items which caused the subgroup mean score differences. In Bridgeport the defendant attempted to establish fairness by

eliminating from the test battery those tests on which there was a significant difference between the scores of blacks and whites. The court, however, was not swayed by this approach for several reasons.

> While this was a commendable measure, the Court is not persuaded that this additional step made the test unbiased. Of necessity the consulting firm's analysis was based on the evaluation of a small number of Blacks, most if not all of whom had demonstrated an ability to perform well on civil service exams. In addition, the firm was never told to eliminate tests that had a discriminatory impact on Hispanics (p. 755).

In Pegues test fairness was established on the basis of a differential validity analysis. This analysis revealed the validities of the test for black and white subsamples were of a similar magnitude (i.e., subgroup validity coefficients were not significantly different). Moreover the court ruled that "plaintiff's allegation that the wrong model of test fairness was used is without merit" (p. 254).

Two types of test fairness evidence were reported in Walls. One indicated the existence of single-group validity (i.e., unfairness) by demonstrating that seven of the validity coefficients were significant for whites, while only three were significant for blacks. The court's ruling of unfairness was also based on evidence showing that the mean black-white difference in test scores was 6.5 times larger than the mean difference in job performance scores.

Finally, the Cormier court based its determination of test fairness on the Cleary model. This court reviewed the research evidence on single-group and

differential validity and concluded that it was not necessary for the defendant to perform such analyses.

Apparently, most courts are not generally knowledgeable enough to thoroughly understand the issues involved in determining a test's fairness. The cases reviewed showed the courts differed in terms of choosing a test fairness model (e.g., often confusing single-group and differential validity with the concept of fairness) as well as what criteria should be used to demonstrate infeasibility. Until plaintiffs and defendants themselves become able to enlighten the courts, these problems are likely to continue to occur.

Validity Generalization

The Uniform Guidelines specify that a user may rely on validity studies conducted elsewhere under the following conditions: (1) evidence indicates the instrument is valid, (2) the two jobs (the one on which the validation study was conducted and the one for which the selection procedure is to be used) involve the same major work behaviors, and (3) variables which could affect the applicability of the study to the new setting are appropriately taken into account (e.g., performance standards, work methods, representativeness of sample in terms of experience or other relevant factors).

This last requirement is apparently based on the notion that validity is situation specific, and the burden is thus on the defendant to demonstrate the similarity between the job contexts. Principles, however, notes "the current research is showing that the differential effects of numerous variables may not be as great as heretofore assumed" (p. 16). Numerous studies (e.g., Pearlman, Schmidt, & Hunter, 1980; Schmidt, Hunter, Pearlman, & Shane, 1979) have shown that, at least under some circumstances, between-study variance in observed validity coefficients can be explained to a great extent by measurement error.

The issue of validity generalization was addressed in four cases. In three, the courts were guided by the standards set forth in the Uniform Guidelines that additionally require the defendant to demonstrate the similarity between/among job settings. For example, the Dickerson court ruled that a selection test validated in one setting was not "transportable" to a new setting, regardless of whether the original validation study was proper since the defendants presented no evidence demonstrating the similarity between the job settings. The court stated that when one study is used in lieu of another the user "must substantiate in detail job comparability and must show the absence of major differences in contextual variables or sample composition which are likely to affect validity" (p. 1339). The decisions in Ensley Branch and Rivera were also based on whether the defendant met its burden of satisfying the requirements stipulated in the Uniform Guidelines for both content and contextual comparability between/among jobs. These requirements were satisfied in Rivera, but not in Ensley Branch.

In the remaining case (Pegues), the court's ruling was clearly influenced by the current validity generalization research noted in Principles. In upholding the job-relatedness of the examination, the court stated:

> Empirical research has demonstrated that validity is not perceptably changed by differences in location, differences in specific job duties or applicant populations. Plaintiff's allegations that validity is specific to a particular location, a particular set of tasks, and to a specific applicant population...is not true (p. 254).

Apparently, most courts embrace the requirements of the Uniform Guidelines

when examining evidence of validity generalization. Defendants would be wise to consider this when attempting to use validity evidence from other studies to support the validity of their tests (unless, of course, they can persuade the courts of tne credibility of the recent research evidence that validity is not situation-specific).

Discussion

This article reviewed 12 court decisions since 1978 for the purpose of assessing the implications of the latest professional and legal guidelines for the adjudication of cases involving the criterion-related validity of paper-and-pencil tests. In general, most courts very carefully scrutinized these validation studies, and were often found to be quite critical in their evaluations. In fact, the criterion-related validity of the predictor was upheld in only five of the 12 cases reported.

The standards imposed on validation studies varied from court-to-court. While most written decisions made reference to the legal and professional guidelines, specific requirements contained in them have not been uniformly applied. The areas of major disagreement between the information contained in the courts' decisions and the legal and professional guidelines concerned the following: (1) whether or not it is important to base selection instruments on systematic developmental procedures (i.e., a job analysis), (2) whether the plaintiff or defendant is responsible for an alternatives search, (3) the degree to which the criterion instrument should be scrutinized, (4) the importance of statistical power issues when considering the statistical and practical significance of validity results, (5) the evidence necessary for establishing test "transportability" for validity generalization purposes (including information on job groupings), (6) whether statistical corrections for

attenuation and range restriction are appropriate, and (7) whether test fairness is a legitimate issue, and if it is, which model should be used to assess fairness.

At least two of these areas of disagreement are exacerbated by the fact that the information in the Uniform Guidelines and Principles appear to be at odds. These areas are test fairness and validity generalization. The professional literature on which the recommendations in the Principles on these two issues is based supports the idea that tests do not predict differentially for majority and minority subgroups, and that validity evidence may be transportable across job contexts.

Most of the courts which have directly addressed these two issues have sided with the Uniform Guidelines. These legal guidelines explicitly require the assessment of test fairness (where feasible), as well as evidence of similarity between/among jobs with regard to both content and context in order to justify the transportability of a test from one setting to another.

The lack of consistency and uniformity among the courts about the issues addressed in this paper is also due in part to the varying strategies employed in the adjudication of the cases. Two separate adjudication strategies appear to exist—"judge as expert" and "psychologist as expert." The "judge as expert" strategy is followed by courts which apparently have little desire to "umpire a battle of experts." These courts focus their attention on the content of examinations. In some instances, judgments are influenced by the predictor's face validity; in others, by its content validity. Such courts place a heavy burden on the defendant to present strong statistical evidence of validity if dissatisfied with either the procedures used to develop or select the exam, the content of the examination, or the exam's "track record."

The courts employing the "psychologist as expert" strategy base their decisions on the testimony given by expert witnesses (usually psychologists). This testimony generally centers on the witnesses' opinions as to what constitutes an acceptable validation study. The Uniform Guidelines, Principles and/or recent findings in selection research are often cited in defense of their positions. In most instances (not surprisingly), the testimony provided by the plaintiff's and defendant's witnesses conflict, and the court is placed in the difficult position of having to resolve the disagreement to reach a decision.

In most instances it was difficult to accurately discern how the courts resolved these conflicts. Sometimes, conflicts were resolved on the basis of the witnesses' credibility. For instance, the Virginia court ruled in favor of the defendant because it did not believe that the plaintiff's witness "testified as a disinterested expert. His every aspect, demeanor, expression, inflection, and response was characterized by partisanship for the side which had hired him" (p. 1100).

The development of an easily understandable document which would reflect a consensus among professionals about the various technical issues would provide the courts with the information necessary to resolve disagreements among expert witnesses. The use of such a document would surely increase the uniformity among court decisions. Hopefully, the Joint Technical Standards, when published, will serve that purpose.

Recommendations

The following recommendations are offered as a guide to practioners who are faced with the task of performing and legally defending a criterion-related validation study. These recommendations represent, in part, a restatement and elaboration of suggestions previously made in this paper.

Conducting a Validation Study

The first step facing practitioners is to assess the existing selection procedure. If its use has not resulted in adverse impact against any protected group, validation would not legally be required (although it would certainly be desirable). If there is evidence of adverse impact, two alternatives exist: directly remove the adverse impact or demonstrate the job relatedness of the selection procedure (i.e., validate it).

The easiest way to remove adverse impact would be to equate the selection ratios of the various protected groups by altering cutoff scores. While such an approach would appear to be condoned by the Uniform Guidelines, the courts have not been uniform in their acceptance of this practice.

Because of this lack of uniformity, and because of other considerations (e.g., the detrimental impact of lowered cutoff scores on validity and utility), we do not recommend the direct removal of adverse impact. Rather, we advocate the second alternative, viz., the demonstration of validity. Using this approach, the employer has a choice of validating the existing instrument or developing and validating a new one.

A job analysis should be performed prior to making this decision. From a legal perspective, it is important that the job analysis identify job groupings, the specific knowledge, skills, and abilities (KSAs) needed for each job(s), as well as indicate their relative importance at entry. If more than one job group is identified, validation should be conducted separately for each group. We further recommend that the current selection procedure be validated only if the results of the job analysis indicate that it assesses these important KSAs. If not, a new selection procedure should be chosen or developed prior to validation.

The next decision that has to be made is whether to conduct a local validation study (either in-house or co-operative) or to rely on validity generalization data. The latter approach would be unacceptable to most courts unless it could be established that the job(s) for which the instrument was originally validated is/are similar to the present job(s) in terms of both content and context. If not, a local validation study should be conducted.

At some point prior to conducting a local validation study, the employer should investigate the possibility that an alternative procedure exists which may be as valid, and yet have less adverse impact than the current or proposed procedure. We recommend this "search" despite the case law which appears to place the burden of a search on the plaintiff. Our reasoning is that the search will enable the employer to collect data on possible alternatives which may be suggested by the plaintiff in court. Moreover, if a suitable alternative were found at this time, an employer would have the opportunity to implement it prior to validation, and thus, present a more favorable defense posture.

The next step would be to develop an adequate criterion measure. Most courts require that the criterion instrument be derived from job analyses data. A new criterion instrument should be developed if an existing one has not been based on a thorough job analysis. In any event, whether using a new or existing criterion instrument, it is important to differentially weight each criterion dimension, and to ensure that dimensional definitions and/or behavioral anchors are clear and specific.

The next step in the empirical validation process should be to determine the particular strategy (i.e., predictive or concurrent) to employ. If feasible, a predictive study should be done. A concurrent study would, however, be acceptable to most courts if the validation sample closely resembled the

typical applicant group with regard to all relevant characteristics (e.g., experience, education, race, sex, etc.). Moreover, the validation sample should be large enough to ensure adequate statistical power.

The test and criterion should be administered under standardized conditions. It is particularly important in a concurrent study to provide some incentive to motivate subjects in the validation sample to perform at their best level on the test. Regarding the collection of criterion data, it is important to provide the raters with adequate instructions, preferably by individually amdinistrating the rating form. The instructions should describe how to complete the rating form, and should stress the need for objectivity. The raters chosen should be familiar with the work performance of the incumbent, and should be given training on how to evaluate performance.

Once the data are collected, we recommend that the psychometric characteristics of the test and criterion be assessed. This would include an assessment of test reliability, test item-characteristics, the reliability, and construct validity of the criterion, as well as the extent of rating errors (e.g., halo, leniency, etc.). We make this recommendation in spite of the fact that these assessments are often not required by the courts.

In determining the relationship between test scores and the criterion measure, it is, of course, important to assess the validity and utility of the predictor. No specific method for assessing utility has been advocated by a majority of the courts. Among those that have been used by defendants are: the coefficient of determination, a comparison of correct selection decisions made with and without the use of the test, the standard error of estimate, a joint consideration of the selection ratio and the risk associated with the selection of false positives, and the Brogden model.

Finally, we recommend that where feasible, the examination's fairness for various protected groups be assessed. The Cleary model is the preferred method.

Defending a Criterion-Related Validation Study in Court

A low validity coefficient (which may or may not be statistically significant) will generally trigger a more thorough scrutiny of the validation study by the courts. The following suggestions are intended to assist employers in the preparation of their legal defense should this contingency arise.

The goal of the employer should be to convince the court that the low coefficient was the result of factors other than the test itself. This approach would be warranted only if the selection instrument were carefully developed on the basis of job analysis information. The argument for the defense could then be augmented by information on the test's content validity.

To persuade the court that the low validity coefficient may be an underestimation of the test's job-relatedness, the defense should discuss the issues of statistical power, range restriction in the predictor and criterion, and attenuation due to criterion unreliability. When statistically correcting the validity coefficient for range restriction and attenuation, the employer must be prepared to provide evidence that the assumptions of these statistical tests are met.

Finally, the importance of credible expert witnesses cannot be ignored. If all else fails to favorably persuade the court, the testimony of expert witnesses may be enough to influence the court's decision more positively, especially the decisions of those courts that employ the "psychologist as expert" adjudication strategy. The additional costs of such experts may well be worth the investment.

This last point begs at the question of professional ethics, which should

not be ignored. "Credibility" and ethics go hand-in-hand. Whatever services are offer to plaintiffs or defendants by professionals must be "congruent with proper and ethical practice in our field.... All research, be it basic or applied, is done within constraints. It is up to the individual I/O psychologist to behave ethically within the constraints of his/her situation" (Michael Kavanaugh's letter to the Industrial-Organizational Psychologist, p. 65). No where is this more important than in a court of justice.

References

Ass'n Against Discrimination v. City of Bridgeport, 454 F. Supp. 751 (D. Conn. 1978).

Bemis, S.E. (1968). Occupational validity of the general aptitude test battery. Journal of Applied Psychology, 52, 240-244.

Bobko, P. (1983). An analysis of correlations corrected for attenuation and range restriction. Journal of Applied Psychology, 68, 584-589.

Boehm, V.R. (1977). Negro-white differences in validity of employment and training selection procedures: Summary of research evidence. Journal of Applied Psychology, 62, 146-154.

Bushey v. New York State Civil Service Commission, 571 F Supp. 1562 (N.D. New York 1983).

Cascio, W.F. (1982). Applied psychology in personnel management (2nd ed.). Reston, Virginia: Reston.

Commonwealth of Pennsylvania v. O'Neill, 465 F. Supp. 451 (E.D. Penn. 1979).

Cormier v. PPG Industries, 26 FEP Cases 652 (W.D. Louisiana, 1981).

Darlington, R.B. and Stauffer, F.G. (1966). A method for choosing a cutting point on a test. Journal of Applied Psychology 50, 229-231.

Dickerson v. U.S. Steel Corp., 472 F. Supp. 1304 (E.D. Penn. 1978).

Dreher, G.F. and Sackett, P.R. (1983). Perspectives employee staffing and selection. Homewood, Ill.: Richard D. Irwin.

Ensley Branch v. Seibels, 22 FEP Cases 1207 (Fifth Circuit, 1980).

Guardians Association v. Civil Service Commission of the City of New York, 630 F.2d 79 (Second Circuit 1980).

Gutenberg, R.L., Arvey, R.D., Osburn, H.G. and Jeanneret, P.R. (1983). Moderating effects of decision-making/information processing job dimensions on test validities. Journal of Applied Psychology, 68, 602-608.

Ghiselli, E.E. (1973). The validity of aptitude tests in personnel selection. Personnel Psychology, 26, 461-477.

Hunter, J.E., Schmidt, F.L. and Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis.

I.M.A.G.E. v. Bailor, 518 F. Supp. 800 (N.D. Cal. 1981).

Industrial-Organizational Psychologist. (1984). Society for Industrial and Organizational Psychology, Inc., Division 14 of the American Psychological Association, 21(3), May.

Landy, F.J. and Farr, J.L. (1980). Performance rating. Psychological Bulletin, 87, 72-107.

Pearlman, K. (1980). Job families: A review and discussion of their implications for personnel selection. Psychological Bulletin, 87, 1-28.

Pearlman, K., Schmidt, F.L. and Hunter, J.E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical operations. Journal of Applied Psychology, 65, 373-406.

Pegues v. Mississippi State Employment Service, 488 F. Supp. 239 (N.D. Miss. 1980).

Principles for the validation of selection procedures: Second Edition. (1980). Division of Industrial-Organizational Psychology, American Psychological Association.

Reilly, R.R. and Chao, G.T. (1982). Validity and fairness of some alternative employee selection procedures. Personnel Psychology, 35, 1-62.

Rivera v. City of Wichita Falls, 665 F.2d (Fifth Circuit 1982).

Rorer, L.G., Hoffman, P.J., LaForge, G.E., and Hsieh, K.C. (1976). Optimimum cutting scores to discriminate groups of unequal size and variance. Journal of Applied Psychology, 50, 153-164.

Sackett, P.R. and Wade, B.E. (1983). On the feasibility of criterion-related validity: The effects of range restriction assumptions on needed sample size. Journal of Applied Psychology, 68, 374-381.

Schmidt, F.L. and Hunter, J.E. (1980). The future of criterion-related Validity. Personnel Psychology, 33, 41-60.

Schmidt, F.L. and Hunter, J.E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529-540.

Schmidt, F.L., Hunter, J.E., Pearlman, K. and Shane, G.S. (1979). Further tests of the Schmidt-Hunter Bayesian validity generalization procedure. Personnel Psychology, 32, 257-281.

Schmidt, F.L., Hunter, J.E., Croll, P.R., and McKenzie, R.C. (1983). Estimation of employment test validities by expert judgement. Journal of Applied Psychology, 68, 590-601.

Schmidt, F.L., Hunter, J.E. and Urry, V.W. (1976). Statistical power in criterion-related validity studies. Journal of Applied Psychology, 61, 473-485.

Spool, M.D. (1978). Training programs for observers of behavior: A review. Personnel Psychology, 31, 853-888.

Uniform guidelines on employee selection procedures (1978). Federal Register, 43, 38290-38315.

United States v. City of Chicago, 573 F.2d 416 (Seventh Circuit 1978).

United States v. Commonwealth of Virginia, 454 F. Supp. 1077 (E.D. Virginia 1978).

Walls v. Mississippi State Department of Public Welfare, 542 F. Supp. 281 (N.D. Miss. 1982).

Footnotes

[1]Alternatively, the situation could arise where the use of a less valid test (with less adverse impact) but with greater utility might be considered a "suitable" alternative by the courts. Obviously, there are always less valid tests available to employers, and many of these may potentially have equal or greater utility with similar or less adverse impact.

If equal utility rather than equal validity should ever become the focus of an alternative search, this potentially broadens the definition of an "alternative," and may create an even larger burden for whomever is judged responsible for the search. However, if utility is viewed as the sine qua non of testing (Principles), maybe this is the correct way for the courts to proceed.

[2]One other approach for lowering adverse impact that could be considered a variant of adjusting cutoff scores is the altering of predictor scores. This approach was used in Bushey. Here, the employer attempted to directly remove adverse impact by adding 250 points to the test scores of each minority candidate. Although suit in Bushey was brought on both Title VII and constitutional grounds, the court ruled against the employer only on the basis of its determination that this practice violated the Fourteenth Amendment rights of non-minority candidates.